

Models with time-varying predictors for meningitis in Navrongo, Ghana

Yolanda Hagar, Mary Hayden, Abudulai Adams Forgor, Patricia Akweongo,
Abraham Hodgson, Christine Wiedinmyer, Vanja Dukic *

Abstract

The “meningitis belt” is a region in sub-Saharan Africa where annual outbreaks of meningitis occur, with large epidemics observed cyclically. While we know that meningitis is heavily dependent on seasonal trends (in particular, weather), the exact pathways for contracting the disease are not fully understood and warrant further investigation. This manuscript examines meningitis trends in the context of survival analysis, quantifying underlying seasonal patterns in meningitis rates through the hazard rate for the population of Navrongo, Ghana. We compare three candidate models: the commonly used Poisson generalized linear model, the Bayesian multi-resolution hazard model, and the Poisson generalized additive model. We compare the accuracy and robustness of the models through the bias, RMSE, and the standard deviation. We provide a detailed case study of meningitis patterns for data collected in Navrongo, Ghana.

Key Words: survival analysis, hazard rate, count data, multi-resolution hazard, changing at-risk population, time-varying covariates

*Yolanda Hagar is a postdoctoral researcher in applied mathematics at the University of Colorado at Boulder. Mary Hayden and Christine Wiedinmyer are scientists at the National Center of Atmospheric Research in Boulder, Colorado. Abudulai Adams Forgor is the medical superintendent at the War Memorial Hospital, Ghana Health Service in Navrongo, Ghana. Patricia Akweongo is a senior lecturer at the Department of Health Policy, Planning and Management, School of Public Health at the University of Ghana in Accra, Ghana. Abraham Hodgson is director of the Research and Development Division of Ghana Health Services, Accra, Ghana. Vanja Dukic is a professor in applied mathematics at the University of Colorado at Boulder.
Vanja Dukic (email: Vanja.Dukic@colorado.edu)

1 Introduction

Within the “meningitis belt”, a region in sub-Saharan Africa stretching from Ethiopia to Senegal (also known as the Sahel), outbreaks of meningococcal meningitis happen annually during the dry season, with large epidemics occurring every 2 to 10 years [34, 20]. The dynamics of transmission of meningitis in the Sahel are not fully explained, as the causes are likely due to a variety of factors, including socio-economic status, migration, environmental circumstances, and the introduction of new strains of the disease. While meningitis can be controlled through vaccination, this strategy is dependent on the identification of populations that are at risk and periods of time when transmission is most likely. Meningitis is a serious illness (with approximately 1 in 10 cases being fatal [41]), however, the actual number of diagnosed meningitis cases can be small when compared to the number of people at risk (ranging from 10 to 1000 infections per 100,000 people [22]), making inference about the wide range of contributing factors difficult.

Meningitis is highly dependent on seasonal trends, with larger epidemics occurring cyclically over multiple year time-spans. Several studies support the idea that weather conditions greatly impact the transmission of meningitis, with higher incidence rates occurring between December and May when the weather is dry [1, 5, 9, 11, 20, 21, 34, 39, 40, 38, 48, 52, 50, 51, 53, 56]. However, even with reliable weather forecasting, patterns in yearly and cyclic meningitis rates cannot be entirely explained by changes in climate, as there are many other temporal factors that co-vary with weather patterns (such as changes in social activities, patterns of migration, and variations in cooking methods). These other trends can be more difficult to measure but may have an equally significant impact on the transmission of meningitis. Therefore, statistical models quantifying trends in meningitis outbreaks must be able to account for seasonal covariates as well as yearly trends in the disease due to latent variables.

Many statistical approaches have been used to quantify meningitis rates in other areas of the meningitis belt. [4] created a Bayesian network to model how meningitis incidence rates change based on dependent relationships among different districts in Niger. This model is a graphical tool that showed the probabilistic relationships of contracting meningitis among different regions. [8] used a wavelet method to compare and contrast different time series of counts for various countries in the meningitis belt, examining both yearly trends and cyclical epidemics. Neither of these two analyses incorporated weather covariates or other predictors in the modeling of meningitis counts, but they did characterize general trends and peaks in the disease over many decades. Other types of modeling used a generalized linear model approach, assuming a negative binomial distribution for meningitis counts [18]. Work by [49] compared different types of models, including dynamic regression models (using a Bayesian framework), and a three-state Markov chain model estimating the probability of changing between different incidence risk levels. In [12], the authors examined individual and group-level data, and modeled the risks based on both individual outcomes and spatially related measures using a Poisson cluster point process. In a similar vein, [2] discretized weekly incidence rates into latent states based on different epidemiological thresholds, modeling the probability of transitioning between states using a multi-state Markov model, accounting for climate variables and dependencies based on a variety of factors. These methods are useful, and account for both changes in climate and spatial information over time. However, the data for the analyses mentioned above mostly come from Niger, which experiences large epidemics every 8 to 12 years and generally has a higher incidence rate than Ghana, which has not had a cyclic epidemic pattern since the 1960’s [8]. In this manuscript, we examine how three similar models behave under circumstances with low event counts and sparse amounts of information, as is the case with the meningitis counts for our data from Navrongo, Ghana and the surrounding Kassena-Nankana District.

Meningitis Data

The data we examine in this manuscript are 364 meningitis cases, aggregated as monthly counts over an 11-year period from 1998 to 2008 in Navrongo and the surrounding Kassena-Nankana (K-N) District in Northern Ghana. Weather measurements are also reported, including daily dust status, number of sunshine hours, maximum and minimum daily temperature, relative humidity, rain quantity, and wind speed. Because the meningitis counts are monthly aggregated figures, the weather variables are provided as monthly averages or percentages. (See [37] and [16] for details on the data set.)

This particular data set from Navrongo was analyzed by [16], using a generalized additive model (GAM; [27]) assuming a Poisson distribution on the meningitis counts, allowing for a smooth function to account for underlying baseline trends in the seasonality of meningitis while estimating the impact of changes in weather. The authors explore optimal models for describing patterns in meningitis by using biological rationale, reduction of collinearity among variables, and the Akaike Information criteria (AIC; [3]).

We expand on this work, and present our examination of this data in the context of survival analysis, quantifying underlying seasonal patterns in meningitis rates through the hazard rate while estimating the impact of the time-varying effects of weather. Our goal is to compare differences in estimation among three suitable models: 1) The ubiquitous Poisson generalized linear model (GLM; [43]), also referred to in some literature as the log-linear model; 2) The Bayesian multi-resolution hazard model (MRH; [6, 7, 10, 13, 15, 14]), which has a likelihood function identical to the first model, but uses a tree-like prior that allows for sharing of information across time periods, making it an ideal method for estimation when the number of observed failures are low; and 3) The Poisson generalized additive model (GAM; [27]), which is similar to the previous two models, but the hazard rate is estimated with a smooth function instead of a piecewise-constant function. We employ the well-established link between the log-linear model and the piecewise exponential hazard model ([28, 33]), as time-varying covariates and a changing (i.e. not monotonically decreasing) at-risk population are easily incorporated into the piecewise exponential model.

This paper is organized as follows: In Section 2, we compare theoretical differences among the three different models (MRH, GLM, and GAM) and examine how they relate to the piecewise exponential model, and in Section 3 we examine the variance for the Poisson GLM and MRH model. In Section 4, we compare the robustness and accuracy of the three models through simulations, and in Section 5, we examine the Navrongo data set. We conclude in Section 6 with a discussion.

2 Hazard Models for Meningitis

Hazard modeling is commonly associated with survival analysis. However, the hazard rate function, defined as $h(t) = \lim_{\Delta \rightarrow 0} P(t \leq T < t + \Delta \mid T \geq t) / \Delta = f(t) / S(t)$ (where $S(t)$ is the survival function and $f(t) = -S'(t)$), is also a convenient method for characterizing a process over time. In this manuscript, the hazard rate is used to describe counts of events (such as the number of meningitis cases) over the course of a year for an inhomogeneous Poisson process. While there are a plethora of methods for estimation of the hazard rate (for examples, see [30, 32, 31]), a simple and convenient method for describing the time-varying pattern is with a piecewise constant function. In general theory, this is known as the piecewise exponential (PE) hazards model [19, 47, 29, 28, 33], which easily accommodates time-varying covariates (such as monthly weather measurements) as well as a changing number of subjects at risk (such as the increase in the population of Navrongo between 1998 and 2008). We are interested in examining how this model (represented through the Poisson generalized linear model [43]) compares to the modern, Bayesian MRH model, which can also be expressed as a PE model but has a prior structure that lends itself to estimation in the instance of small observed failures. We contrast these two models to a Poisson GAM, as it estimates the

hazard rate with a smooth function while still accommodating time-varying weather covariates and a changing at-risk population.

In the following subsections, we compare the likelihood formulation of the Poisson GLM, the MRH model, and the Poisson GAM. All models have an underlying function that accounts for seasonal baseline meningitis risk, accounting for trends that cannot be fully explained by weather covariates. In addition, all three models can accommodate time-varying weather variables, as well as a changing population at risk.

Piecewise Constant Hazard Models

Below, we review and define notation for PE models and then contrast them to our models in the following sections. As we will show, both the Poisson GAM and the MRH model have likelihoods identical to the piecewise exponential model, with a similar form for the Poisson GAM.

In the context of survival analysis, a “failure” is defined as the contraction of meningitis, and the “survival time” represents the time from the beginning of the seasonal year (i.e. June) until contraction of meningitis. Subjects are censored due to death or because the 11-year observation period is over. Let T_i represent the survival time of the i^{th} individual, and let δ_i be the censoring indicator, which equals ‘1’ if the failure is observed and ‘0’ otherwise. For a piecewise approximation of the hazard rate, the study period $[t_0, t_J]$ is divided into J intervals, $(t_{j-1}, t_j), j = 1, \dots, J$, and a constant baseline hazard rate is estimated within each time interval (or “bin”). Let τ_{ij} denote the time subject i spent in interval j , such that

$$\tau_{ij} = \begin{cases} 0 & \text{if } T_i < t_{j-1} \\ T_i - t_{j-1} & \text{if } T_i \in [t_{j-1}, t_j] \\ t_j - t_{j-1} & \text{if } T_i > t_j. \end{cases} \quad (1)$$

Similarly, we can define δ_{ij} such that it equals ‘1’ if individual i had an observed failure in bin j , and is ‘0’ otherwise, noting that $\delta_i = \sum_{j=1}^J \delta_{ij}$. The baseline cumulative hazard function is defined as $H_0(T_i) = \sum_{j=1}^J \tau_{ij} \lambda_j$, where λ_j is the constant baseline hazard rate in bin j . Covariates (both time-varying and time-independent) can easily be incorporated into the model by defining the cumulative hazard as

$$H(T_i \mid X_i, Z_i, \vec{\beta}, \vec{\alpha}, \vec{\lambda}) = \left(\sum_{j=1}^J \tau_{ij} \lambda_j \exp\{Z'_{ij} \vec{\alpha}\} \right) \exp\{X'_i \vec{\beta}\},$$

where for subject i , X_i is a $1 \times p$ vector of p time-independent (e.g. as gender) covariate measures, Z_{ij} is a $1 \times q$ vector of q time-varying (e.g. monthly temperature) covariate measures in bin j , and $\vec{\beta}$ and $\vec{\alpha}$ are, respectively, the effects of the covariates on the time to survival. (Note that this formulation assumes that the time-varying covariates change only at the boundaries of each bin, which is a valid assumption for our data but may require more advanced techniques with other data sets.)

Likelihood Function

Using the notation above, the log-likelihood function for subject i can be written as:

$$\begin{aligned} \log(L_i | T_i, \delta_i, X_i, Z_i, \vec{\beta}, \vec{\alpha}, \vec{\lambda}) &= \delta_i \log h(T_i | X_i, Z_i, \vec{\beta}, \vec{\alpha}, \vec{\lambda}) - H(T_i | X_i, Z_i, \vec{\beta}, \vec{\alpha}, \vec{\lambda}) \\ &= \sum_{j=1}^J \left\{ \delta_{ij} \left(\log(\tau_{ij} \lambda_j) + X_i' \vec{\beta} + Z_{ij}' \vec{\alpha} \right) - \tau_{ij} \lambda_j \exp\{X_i' \vec{\beta} + Z_{ij}' \vec{\alpha}\} \right\}. \end{aligned}$$

The standard piecewise exponential model can be easily modified to accommodate grouped observations (i.e. multiple failures at the same time point) as well as a changing population of subjects at risk over time. In this instance, i denotes a subject *group*, such as the subjects who contract meningitis within a year i . We define Δ_{ij} as the number of subjects in group i who failed in bin j and Γ_{ij} as the number of subjects right-censored in bin j . In this instance, the log-likelihood contribution for subjects from group i who fail in bin j (either as an observed or censored failures) can be rewritten as:

$$\begin{aligned} \log(L_{ij} | T_i, X_i, \mathbf{Z}_i, \vec{\beta}, \vec{\alpha}, \vec{\lambda}) \\ = \Delta_{ij} \left(\log(\tau_{ij} \lambda_j) + X_i' \vec{\beta} + Z_{ij}' \vec{\alpha} \right) - (\Delta_{ij} + \Gamma_{ij}) \sum_{\ell=1}^j \tau_{i\ell} \lambda_\ell \exp\{X_i' \vec{\beta} + Z_{i\ell}' \vec{\alpha}\}, \end{aligned} \quad (2)$$

noting that $\tau_{i\ell}$ is equal to zero for all bins past the failure bin j . While piecewise exponential models do not require that all time bins be the same width, in our analysis we divide the study time into equal lengths since our data is reported monthly. If the bin width is denoted as ω (i.e. $\omega = t_j - t_{j-1}, j = 1, \dots, J$), the cumulative hazard in equation (2) for all subjects at risk in group i across all J bins can be rewritten as:

$$\begin{aligned} &\sum_{j=1}^J (\Delta_{ij} + \Gamma_{ij}) \sum_{\ell=1}^j \tau_{i\ell} \lambda_\ell \exp\{X_i' \vec{\beta} + Z_{i\ell}' \vec{\alpha}\} \\ &= \exp\{X_i' \vec{\beta}\} \sum_{j=1}^J \omega \lambda_j \exp\{Z_{ij}' \vec{\alpha}\} \left(\frac{\tau_{ij}}{\omega} (\Delta_{ij} + \Gamma_{ij}) + \sum_{\ell=j+1}^J (\Delta_{i\ell} + \Gamma_{i\ell}) \right) \\ &= \exp\{X_i' \vec{\beta}\} \sum_{j=1}^J \omega \lambda_j \exp\{Z_{ij}' \vec{\alpha}\} \left(\frac{\tau_{ij}}{\omega} (\Delta_{ij} + \Gamma_{ij}) + \sum_{\ell=j}^J (\Delta_{i\ell} + \Gamma_{i\ell}) - (\Delta_{ij} + \Gamma_{ij}) \right) \\ &= \exp\{X_i' \vec{\beta}\} \sum_{j=1}^J \omega \lambda_j \exp\{Z_{ij}' \vec{\alpha}\} \Phi_{ij}, \end{aligned} \quad (3)$$

with $\Phi_{ij} = \frac{\tau_{ij}}{\omega} (\Delta_{ij} + \Gamma_{ij}) + N_{ij} - (\Delta_{ij} + \Gamma_{ij})$. Note that $N_{ij} = \sum_{\ell=j}^J (\Delta_{i\ell} + \Gamma_{i\ell})$ represents the number of subjects at risk in group i at the beginning of bin j , and that this formulation of the model allows for new subjects to become at-risk over the course of the observation period.

The log-likelihood for group i can therefore be written as:

$$\log(L_i | T_i, X_i, \mathbf{Z}_i, \vec{\beta}, \vec{\alpha}, \vec{\lambda}) = \sum_{j=1}^J \Delta_{ij} \left(\log(\tau_{ij} \lambda_j) + X_i' \vec{\beta} + Z_{ij}' \vec{\alpha} \right) - \exp\{X_i' \vec{\beta}\} \sum_{j=1}^J \omega \lambda_j \exp\{Z_{ij}' \vec{\alpha}\} \Phi_{ij}. \quad (4)$$

In the next sections, we will show that the Poisson GLM and the MRH model have likelihood functions equivalent to the one in equation (4), and that the Poisson GAM is similar, with a smooth approximation for the baseline hazard rate.

2.1 Poisson GLM model

As shown in [28] and [33], the piecewise exponential model is equivalent to a Poisson regression model, which assumes the number of subjects who contract meningitis within each bin follows a Poisson distribution. Use of the Poisson model in a survival analysis context assumes that the number of failures within each bin of time is independent. While this assumption cannot possibly hold in the standard survival analysis setting (as failure of a subject in an early bin is directly related to the number of subjects that can fail in later bins), in this particular context the at-risk population is constantly changing, making the assumption of independence between bins more viable. In addition, as pointed out in [33] and [28], this method can be used to derive an equivalent log-likelihood function, as we do below.

Likelihood function

Let y_{ij} represent the number of observed failures from group i in bin j . Then, y_{ij} comes from a Poisson distribution with mean μ_{ij} , such that

$$\mu_{ij} = W_{ij} \lambda_j \exp\{X_i' \vec{\beta} + Z_{ij}' \vec{\alpha}\},$$

where W_{ij} is the total time spent by all subjects from group i in bin j , and can be re-expressed as

$$W_{ij} = (\Delta_{ij} + \Gamma_{ij}) \tau_{ij} + (N_{ij} - (\Delta_{ij} + \Gamma_{ij})) \omega = \omega \Phi_{ij}.$$

In the expression above, $(\Delta_{ij} + \Gamma_{ij}) \tau_{ij}$ represents the length of time the subjects who failed in bin j (censored or not) spent in bin j , and $(N_{ij} - (\Delta_{ij} + \Gamma_{ij})) \omega$ represents the amount of time subjects who survived beyond bin j spent in bin j (with $\tau_{ij} < \omega$ for the subjects who failed in bin j). The log-linear model (i.e. the Poisson GLM) can be expressed as

$$\log \mu_{ij} = \log W_{ij} + \log \lambda_j + X_i' \vec{\beta} + Z_{ij}' \vec{\alpha}, \quad (5)$$

where λ_j represents the baseline seasonal trends in meningitis rates (i.e. the baseline hazard rate).

In this instance, the log-likelihood for group i over all J bins written as:

$$\begin{aligned} \log L_{i,Poiss}(y_i | X_i, Z_i, \vec{\beta}, \vec{\alpha}, \vec{\lambda}) &= \sum_{j=1}^J \{y_{ij} \log \mu_{ij} - \log(y_{ij}!) - \mu_{ij}\} \\ &\propto \sum_{j=1}^J y_{ij} \left(\log \lambda_j + X'_i \vec{\beta} + Z'_{ij} \vec{\alpha} \right) - \exp\{X'_i \vec{\beta}\} \sum_{j=1}^J \omega \lambda_j \exp\{Z'_{ij} \vec{\alpha}\} \Phi_{ij} \end{aligned} \quad (6)$$

Equation (6) is equivalent to equation (2) by noting that the number observed failures in each bin, y_{ij} , is equal to Δ_{ij} . Estimation of the parameters is done through iteratively weighted least squares (IWLS) using standard GLM methods [43], with an offset value equal to $\log(W_{ij})$ for each observation.

2.2 Multi-resolution Hazard Model

The multi-resolution hazard (MRH) model, shown in [6, 7, 15, 13, 14, 10] and used in [23] is a Bayesian piecewise-constant exponential hazard model that is based on a tree-like, wavelet-based multi-resolution prior on the hazard function, which allows for scalability and consistency across different time scales (i.e minutes, weeks, years, etc). The MRH model is closely related to the Polya tree [17, 36]. The Polya-tree prior is an infinite, recursive, dyadic partitioning of a measurable space Ω , although in practice this process is terminated at a finite level M , resulting in “finite” Polya trees. Polya trees have been adapted for modeling survival data in a number of ways (for example, see [42, 26, 25, 58]), including stratified Polya tree priors ([57]), Polya trees with fused bins ([55]), and smoothed Polya tree priors [44]). The MRH prior is a type of Polya tree; it uses a fixed, pre-specified partition, and controls the hazard level within each bin through a multi-resolution parameterization. This parameterization allows parameters to differ across bins and levels of the tree in such a way that the marginal priors at higher levels of the tree are the same, regardless of the priors at lower levels of the tree.

In the MRH model, the hazard rate is approximated in a piecewise fashion with a set of corresponding hazard increments, $d_j, j = 1, \dots, J$, where each d_j represents the aggregated hazard rate over the j^{th} bin. (i.e. $d_j = \int_{t_{j-1}}^{t_j} h(s)ds$ in standard survival notation). We assume that $J = 2^M$, where $M > 0$ and can be chosen in a variety of ways; for example, using model selection criteria as in [7], or using clinical input, as in [13]. We then define “split parameters” $R_{m,p} = H_{m,2p}/H_{m-1,p}, m = 1, 2, \dots, M-1, p = 1, \dots, 2^m - 1$, where $H_{m,p}$ is recursively defined such that $H_{m,p} \equiv H_{m+1,2p} + H_{m+1,2p+1}$. These split parameters determine the shape of the hazard rate, and are between 0 and 1. A $\text{Gamma}(a, b)$ prior is placed on H , and the prior for $R_{m,p}$ is $\mathcal{Be}(2\gamma_{m,p}k^ma, 2(1 - \gamma_{m,p})k^ma)$. The d_j in the final time resolution are estimated using the split parameters and the estimate of the cumulative hazard.

Likelihood function

Using the notation of the MRH model, we can write the cumulative hazard function as:

$$H_{MRH}(T_i | X_i, Z_i, \vec{\beta}, \vec{\alpha}, H, \vec{R}_{m,p}) = \exp\{X'_i \vec{\beta}\} \sum_{j=1}^J \frac{\tau_{ij}}{\omega} d_j \exp\{Z'_{ij} \vec{\alpha}\}.$$

For the subjects in group i , the log-likelihood function becomes:

$$\log L_{i,MRH}(T_i | X_i, Z_i, \beta, \alpha) = \sum_{j=1}^J \Delta_{ij} \left(\log d_j + X_i' \vec{\beta} + Z_{ij}' \vec{\alpha} \right) - \exp\{X_i' \vec{\beta}\} \sum_{j=1}^J d_j \exp\{Z_{ij}' \vec{\alpha}\} \Phi_{ij}. \quad (7)$$

The log-likelihood in equation (7) is proportional to the log-likelihood in equation (4) (and therefore equivalent to equation (6), with $d_j = \lambda_j \omega$).

Parameter estimation

All parameters in the model are estimated using Markov chain Monte Carlo. This Bayesian approach to modeling the hazard rate is what provides us with the flexibility to estimate hazard rate while remaining computationally feasible, even when the number of observed failures (in this case, the number of observed meningitis cases) is small.

The algorithm consists successive sampling from the following full conditional distributions:

1. The full conditional for H is a gamma density with the shape parameter $a + \sum_{i=1}^n \sum_{j=1}^J \Delta_{ij}$, and rate parameter $b^{-1} + \sum_{i=1}^n \sum_{j=1}^J \exp\left(X_i' \vec{\beta} + Z_{ij}' \vec{\alpha}\right) \mathcal{F}_j \Phi_{ij}$, where

$$\mathcal{F}_j = d_j / H \quad (8)$$

represents the combination of $R_{m,p}$ parameters needed to calculate d_j .

2. The full conditional of each $R_{m,p}$ is:

$$R_{m,p}^{2\gamma_{m,p} k^m a - 1} (1 - R_{m,p})^{2(1 - \gamma_{m,p}) k^m a - 1} \prod_{i=1}^n L_i(T_i | X_i, Z_i, \vec{\beta}, \vec{\alpha}),$$

where the likelihood function is from equation (7).

3. With a $\mathcal{N}(0, \sigma_{\beta_s}^2)$ prior (with a fixed variance) on each baseline covariate effect, β_s ($s = 1, \dots, p$), each has the following full conditional distribution:

$$\pi(\beta_s | \beta_s^-) \propto \prod_{i=1}^n \prod_{j=1}^J \left[\exp\{\Delta_{ij} X_{is} \beta_s\} \times \exp\left\{-d_j \exp(X_i' \vec{\beta} + Z_{ij}' \vec{\alpha}) N_{ij}\right\} \right] \exp\left\{-\frac{\beta_s^2}{2\sigma_{\beta_s}^2}\right\}$$

4. Similarly, with a $\mathcal{N}(0, \sigma_{\alpha_s}^2)$ prior (with a fixed variance) on each time-varying covariate effect, α_s ($s = 1, \dots, q$), each has the following full conditional distribution:

$$\pi(\alpha_s | \alpha_s^-) \propto \prod_{i=1}^n \prod_{j=1}^J \left[\exp\{\Delta_{ij} Z_{ijs} \alpha_s\} \times \exp\left\{-d_j \exp(X_i' \vec{\beta} + Z_{ij}' \vec{\alpha}) N_{ij}\right\} \right] \exp\left\{-\frac{\alpha_s^2}{2\sigma_{\alpha_s}^2}\right\}$$

5. The full conditional distributions for the remaining parameters (a, b, k and $\gamma_{m,p}$) are the same as with the classical MRH and can be seen in Appendix A.

2.3 Poisson Generalized Additive Model

The likelihood functions of the Poisson generalized additive model (GAM) and the piecewise exponential are not identical, but are similar. The main difference is that in the generalized additive model, a smoothed function $g(t)$ is used to represent the baseline hazard rate. This model (initially used to describe meningitis patterns in this data set in [16]) is similar to the log-linear model shown in equation (5):

$$\log \mu_i(t) = \log(W_i(t)) + g(t) + X_i' \vec{\beta} + Z_i(t) \vec{\alpha}, \quad (9)$$

where $g(t)$ is a smooth function that corresponds to the baseline hazard rate, and $Z_i(t)$ represents the value of the time-varying covariate at time t for group i . Note that λ_j corresponds to the average value of $g(t)$ over $t \in [t_{j-1}, t_j]$, $\lambda_j = \frac{1}{\omega} \int_{t_{j-1}}^{t_j} g(t) dt$. Similarly, Z_{ij} represents the average of $Z_i(t)$ over $t \in [t_{j-1}, t_j]$, and W_{ij} is the average amount of time spent by the i^{th} group in bin j , such that $W_{ij} = \int_{t_{j-1}}^{t_j} W_i(t) dt$.

3 Impact of MRH Prior on Estimation Efficiency

When the number of observations increases, Bayesian models tend to perform identically to their frequentist counterparts as the effects of the priors diminishes. Thus, when the observed number of meningitis cases is large (within each interval of time), we expect the Poisson GLM and MRH models to perform identically both in accuracy of parameter estimation as well as efficiency. However, we are interested in comparing the two models when the number of observed failures is small (particularly when compared to the number of people at risk), as is the case with meningitis rates in Navrongo.

Of particular interest is the variance for the J piecewise estimates of the hazard rate. In this section, we compare the variances of the Poisson GLM and MRH models through the negative expected values of the second derivatives of the log-likelihood and the posterior distribution. For notational simplicity, the calculations are performed on a reduced version of the piecewise model: We assume that there are only two bins (i.e. $J = 2$), that subject groups fail in the middle of the bin (i.e. $\tau_{ij} = 0.5$), and that the length of the bin is equal to 1 (i.e. $\omega = 1$) so that $d_j = \lambda_j$. We omit covariates, assume there is only one subject group (i.e. $n = 1$), and omit censoring mid-study. However, these results can be extended to more complex circumstances.

In the instance of two bins and one subject group, omitting the group-specific notation, the log-likelihood function for both the Poisson GLM and MRH model becomes

$$\log L(T | \vec{\lambda}) = (\Delta_1 \log(\lambda_1) - \lambda_1 (N_1 - 0.5\Delta_1)) + (\Delta_2 \log(\lambda_2) - \lambda_2 (N_2 - 0.5\Delta_2)).$$

In this model, $\lambda_1 = d_1 = H \times R_{1,0}$.

Under the Poisson model, the variance for $\hat{\lambda}_1$ and $\hat{\lambda}_2$ can be easily calculated as $\Sigma_{GLM} = [-E(H_{GLM})]^{-1}$, (where Σ_{GLM} is the negative expected value of the second derivative of the log-likelihood function), such that

$$var(\hat{\lambda}_1) = \frac{\lambda_1^2}{N_1(1 - \exp\{-\lambda_1\})}, \quad var(\hat{\lambda}_2) = \frac{\lambda_2^2}{N_1(1 - \exp\{-\lambda_2\})}.$$

Note that $E\Delta_j = R_j P(\text{fail in bin } j) = R_j (1 - \exp(-d_j))$. The Similarly, the negative expected

value of the second derivative of the posterior for the MRH model is equal to:

$$-E(H_{MRH}) = \left[\begin{array}{c} \frac{N_1(1-e^{-d_1})+N_2(1-e^{-d_2})+a-1}{H^2} \\ \phi \end{array} \quad \frac{\phi}{\frac{N_1(1-e^{-d_1})+ak}{R_{1,0}^2} + \frac{N_2(1-e^{-d_2})+ak}{(1-R_{1,0})^2}} \right],$$

where $\phi = N_1(1 - 0.5(1 - e^{-d_1})) - N_2(1 - 0.5(1 - e^{-d_2}))$. Using the Laplace approximation [35], the variance of \hat{d}_1 and \hat{d}_2 (the means of the marginal posterior distributions for d_1 and d_2) are then approximated using a second order multivariate Taylor Expansion, such that

$$\begin{aligned} var(\hat{d}_1) = & \left[1 - \frac{a \{ (1 - R_{1,0})^2 + k(1 - 2(1 - R_{1,0})R_{1,0}) \}}{(1 - R_{1,0})^2} - 2d_1 N_2 \right. \\ & + N_1 \left\{ \frac{6R_{1,0} - 4R_{1,0}^2 - 3}{(1 - R_{1,0})^2} + e^{-d_1}(2 + d_1) \right. \\ & \left. \left. + e^{-d_2} \left(1 + \frac{R_{1,0}^2}{(1 - R_{1,0})^2} - d_1 \right) + 2d_1 \right\} \right] / \\ & \left[\frac{(1 - a + N_1(e^{-d_2} + e^{-d_1} - 2)) \left\{ \frac{ak - N_1(e^{-d_2} - 1)}{(1 - R_{1,0})^2} + \frac{ak - N_1(e^{-d_1} - 1)}{R_{1,0}^2} \right\}}{H^2} \right. \\ & \left. + e^{-2(H+d_1)} \left\{ e^{H+d_1} (N_1 - N_2) - 0.5N_1 e^{2d_1} + 0.5N_1 e^H \right\}^2 \right]. \end{aligned}$$

The variance for \hat{d}_2 is a mirror image of the variance for \hat{d}_1 , with $R_{1,0}$ replaced by $1 - R_{1,0}$, N_1 replaced with N_2 , and so on.

One method to compare the variances of $\hat{\lambda}_j$ and \hat{d}_j when the number of observed failures is small is to examine what happens to the variance as the number of observed failures goes to zero (i.e. as the cumulative hazard H goes to zero). To do this, the Poisson parameters are rewritten such that λ_j is a function of H and ξ_j , the “height” of the hazard rate in bin j divided by the cumulative hazard (i.e. $\lambda_j = H \times \xi_j$). In the two bin model, $\xi_1 = R_{1,0}$, and $\xi_2 = 1 - R_{1,0}$, and in general, $\xi_j = \mathcal{F}_j$ (as seen in equation (8)), where \mathcal{F}_j represents the functions of the $R_{m,p}$ values associated with the hazard rate in bin j .

In examining the limit of the variances as the cumulative hazard goes to zero, both $\lim_{H \rightarrow 0} var(\hat{d}_1) = 0$, and $\lim_{H \rightarrow 0} var(\hat{\lambda}_1) = 0$, so it is important to determine which goes to zero “faster” (i.e. for the same values of H , which variance is smaller). To determine which function is smaller for the same values of H , we take the limit of the derivative of the difference between the two functions as H goes to zero. If we let $g_1(H) = var(\hat{d}_1(H)) - var(\hat{\lambda}_1(H))$, and $g_2(H) = var(\hat{d}_2(H)) - var(\hat{\lambda}_2(H))$ then

$$\lim_{H \rightarrow 0} \left(\frac{\partial}{\partial H} g_1(H) \right) = -\frac{R_{1,0}}{N_1}, \quad \lim_{H \rightarrow 0} \left(\frac{\partial}{\partial H} g_2(H) \right) = -\frac{1 - R_{1,0}}{N_2}.$$

Since the derivatives of the differences are negative, the variance of the constant hazard rate for the MRH model is smaller than the variance for the constant hazard rate of the Poisson GLM as the number of observed failures becomes small.

Figure 1 illustrates exactly how small H needs to be for this to hold true by showing calculated values of $g_1(H)$ for censoring percentages ranging from 75% to almost 100% (corresponding to values of H ranging from 0.30 to 0 using the relationship $S(t) = -\log H(t)$) and different values of $R_{1,0}$. It can be observed that for any value of $R_{1,0}$, once H is small enough (below about $H = 0.2$, or 82% censored), the difference between the two variances is negative (with zero denoted by the solid grey line), meaning that $Var(\hat{d}_1) < Var(\hat{\lambda}_1)$. When H is greater than 0.2 (i.e. the censoring is less than about 82%), the sign of the difference depends on the value of $R_{1,0}$. If $R_{1,0} \geq 0.50$, the sign of the difference is negative for all values of H . However, when $R_{1,0}$ is greater than 0.50 (causing $g_1(H)$ to be negative), the next consecutive bin is associated with $1 - R_{1,0}$, which would be less than 0.5, in which case $Var(\hat{d}_2) > Var(\hat{\lambda}_2)$. With larger values of H , the two models will trade-off neighboring hazard estimates with the larger variances. Denoted on the figure are the average percentage of censored subjects in the simulated data sets ('S') and the percentage of censored subjects in the Navrongo data set ('N'), which are both well below the threshold where the both $g_1(H)$ and $g_2(H)$ become negative. In this regime, the MRH is more efficient with any prior choice for $R_{1,0}$ when compared to the Poisson GLM.

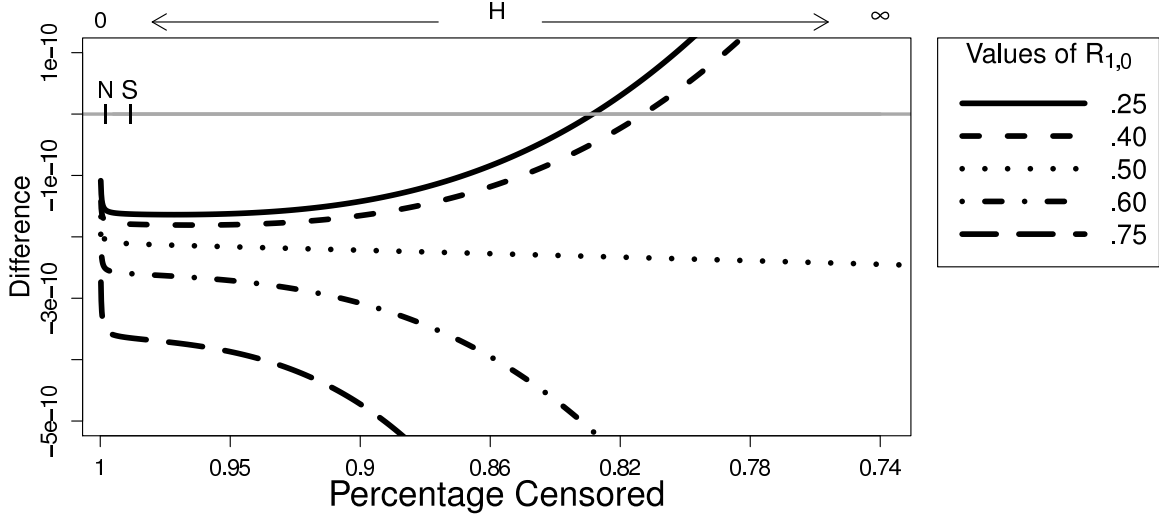


Figure 1: The difference of the variances for the MRH model and the Poisson GLM for the first of two bins as a function of H (i.e. $g_1(H) = var(\hat{d}_1(H)) - var(\hat{\lambda}_1(H))$), with zero denoted by the solid grey line. We see that if H is small enough ($\approx H < 0.20$, censored $\approx 82\%$), then $g_1(H)$ is negative, regardless of the value of $R_{1,0}$. For larger values of H , the function is negative for larger values of $R_{1,0}$. This benefit is minimal in the 2-bin model, however, as a large $R_{1,0}$ in the first bin requires a small $1 - R_{1,0}$ in the second bin. The percentage of censored subjects in the simulated data sets (the average percent censored) is denoted with 'S', and denoted with 'N' for the Navrongo data set. It can be observed that both percentages of censoring ($\approx 98\%$) are well above the threshold where the difference between the variances is always negative.

4 Estimator Assessment Through Simulation

To compare the accuracy and variance of the Poisson GLM, MRH, and the Poisson GAM estimators, we simulated data with weather and failure rate patterns similar to those in the Navrongo, Ghana data set, with about 0.019% of subjects contracting meningitis yearly. The number of deaths in each bin was randomly sampled from a Poisson distribution with $\mu = 120$, and at the beginning of each time bin, the number of subjects at risk was 150,000 (the approximate average population of Navrongo over the 11-year period). The simulated data covered a 10-year period, with meningitis counts and weather variables measured at 16 equally spaced intervals each year. The hazard rate used to generate the survival times had a peak mid-season (representing the peak season for meningitis). In addition, we incorporated two time-varying weather covariates: the monthly average maximum daily temperature and the monthly average humidity at 3pm.

We generated 200 data sets and examined the estimated results from all three models by comparing the mean, the 2.5% and 97.5% quantiles, the within-bin standard deviation, the bias and the RMSE of the 200 estimates. The bias of the hazard rate was calculated for each model at each time point t_j as

$$bias(\hat{h}(t_j)) = 1/200 \sum_{i=1}^{200} (h(t_j) - \hat{h}(t_{ij})), \quad (10)$$

and the RMSE of the hazard rate was calculated at each time point as

$$RMSE(\hat{h}(t_j)) = \sqrt{1/200 \sum_{i=1}^{200} (h(t_j) - \hat{h}(t_{ij}))^2}. \quad (11)$$

Each MRH model was run for 5000 MCMC iterations, with the first 500 iterations burned, and the remaining thinned by 10 to alleviate autocorrelation. Point estimates for the MRH model were calculated as the median of the marginal posterior distribution of each parameter. The Poisson GAM is fitted via a penalized maximum likelihood with $g(t)$ estimated using penalized regression splines. In the simulations, we use a basis dimension equal to 16 and specify a cyclic cubic regression spline to estimate the smooth underlying function, in order to match the value of the function at the end of the season (i.e. June) with the beginning of the season (i.e. July).

The estimated hazard rates and covariate effects for all three models can be seen in Figure 2. The top left graph contrasts the mean of the 200 estimated hazard rates for each of the three models, with the true hazard rate shown in black. All three models perform similarly, although the Poisson GAM seems to overestimate the hazard rate most frequently, and the MRH model seems to underestimate it most frequently. The top right graph shows the 2.5% and 97.5% quantiles of the 200 estimates for each model. The Poisson GLM bounds are the widest throughout most of the study period, with the exception between January and March, when the bounds are very narrow. (Even though the mean of the 200 simulated data sets is close to the true hazard rate, the vast majority of the estimates are very small throughout this period of time, causing the 97.5% quantile to be smaller than the mean.) The 95% intervals graph shows the MRH model has the smallest bounds. Densities of the 200 estimated covariate effects (bottom graphs) show little differences among the three models.

Figure 3 shows the bias and RMSE within each bin (top graphs), and integrated over time (bottom graphs). The within-bin bias shows that the MRH model tends to underestimate the hazard rate, while both the Poisson GLM and GAM tend to overestimate the hazard rate (top, left). However, the MRH model has the smallest absolute bias, as well as the smallest RMSE, integrated

absolute bias, and integrated RMSE across the study period. Alternatively, the Poisson GLM has the largest RMSE, largest integrated absolute bias, and largest integrated RMSE. Similarly, in Figure 4, the within-bin standard deviation of the estimated hazard rates for each model are displayed. As shown in Section 3, the MRH model has a smaller standard deviation when compared to the Poisson GLM. It also appears that the Poisson GAM has a larger standard deviation than the MRH model, but a smaller standard deviation than the Poisson GLM.

5 Analysis of Navrongo Data

Navrongo, located in northern Ghana, experiences two main weather seasons, a wet season lasting from June to October, and then a dry season for the remainder of the year. The population of Navrongo (located in the Kassena-Nankana district) has ranged from 140,000 to 150,000 residents, with approximately 20,000 people living in the city, and the remainder living in rural areas surrounding the city [45, 46]. The data set we examine contains monthly meningitis cases (laboratory confirmed) over an 11-year period from 1998-2008 collected by researchers at the Navrongo Health Research Centre (NHRC). In addition to the meningitis case counts, our data set also contains meteorological data collected by Ghana Meteorological Services, and includes daily dust status, number of sunshine hours, maximum and minimum daily temperature, relative humidity, rain quantity, and wind speed. The data set also contains carbon monoxide (CO) emission estimates, calculated using the Fire INventory model (FINNv1; [54]) based on satellite observations of active fires, and the number of confirmed pneumonia cases each month. The weather data are given as monthly averages or percentages.

While exact population counts were not available for each month, we estimated monthly counts as a linear increase over time, using population information provided in [46, 45] to obtain the approximate number of deaths and births (and hence the number of subjects at risk) each month. Over the 11-year period, a total of 364 meningitis cases (range per year = 0 to 115, SD = 42.6) were documented, and the number of subjects at risk ranged from 141,046 in January, 1998 to 153,236 in December, 2008. The set of covariates we examine in this manuscript was chosen based on previous analyses of this data set, which can be found in [16], where the authors consider multiple models, and discuss the best ones based on biological rationale and through model selection criteria (such as AIC [3]). As such, our model includes the following predictors:

- Maximum daily temperature (degree Celsius, monthly average).
- Minimum daily temperature in previous month (degree Celsius, monthly average).
- Percentage of days of the month with dust (daily measure is a “yes” or “no” binary variable).
- Average relative humidity at 3 pm in previous month (percent).
- Average CO emissions from fires in previous month (grams of CO per day).
- Number of pneumonia cases two months prior.

Certain variables (CO emissions, humidity, and minimum temperature) are thought to have a longer-term effect on the risk of meningitis, and are therefore lagged by one month, while the maximum daily temperature and the amount of dust in the environment are thought to directly impact current health. The pneumonia cases are included as a proxy for respiratory health. (See Table 1 for a summary of the variables used in our analysis.)

In addition to the weather variables, the Poisson GLM included indicators for each of the twelve months (to estimate the constant hazard in each bin), and was implemented using `glm()` with a

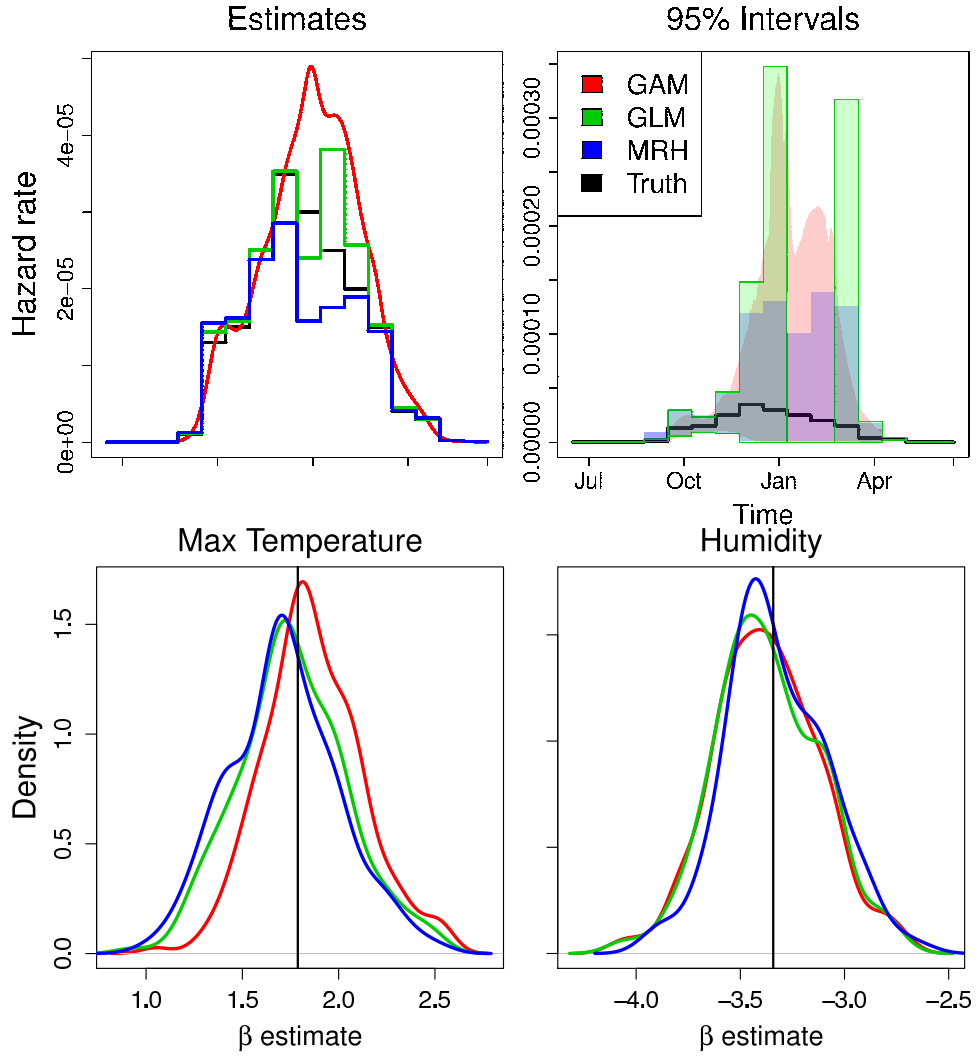


Figure 2: TOP: The estimated hazard rates (left) from the three models, with the true hazard rate superimposed in black. The mean of the 200 estimated hazard rates is shown for each model, and it can be observed that the Poisson MRH model tends to underestimate the true hazard rate, while the other two models tend to overestimate the true hazard rate. On the right, 95% intervals (calculated as the 2.5% and 97.5% quantiles of the 200 estimates) for all three models for the estimated hazard rates. The intervals are narrowest for the Poisson MRH models, and largest for the Poisson GLM. This was expected based on the work shown in Section 3. In addition, we see that while the mean of the Poisson GLM simulations is close to the true hazard rate, the quantiles are all quite small. Both the Poisson GAM And MRH models have quantile bounds that capture the true hazard rate throughout the study. BOTTOM: Densities of the 200 estimated covariate effects for the average max daily temperature (left) and the average daily humidity at 3pm (right) for all three models, with the true parameter values superimposed as a black vertical line. While the Poisson GAM models seem to deviate the most from the other two, there are few differences in the results of the three models. The densities were calculated using the `density()` function in R.

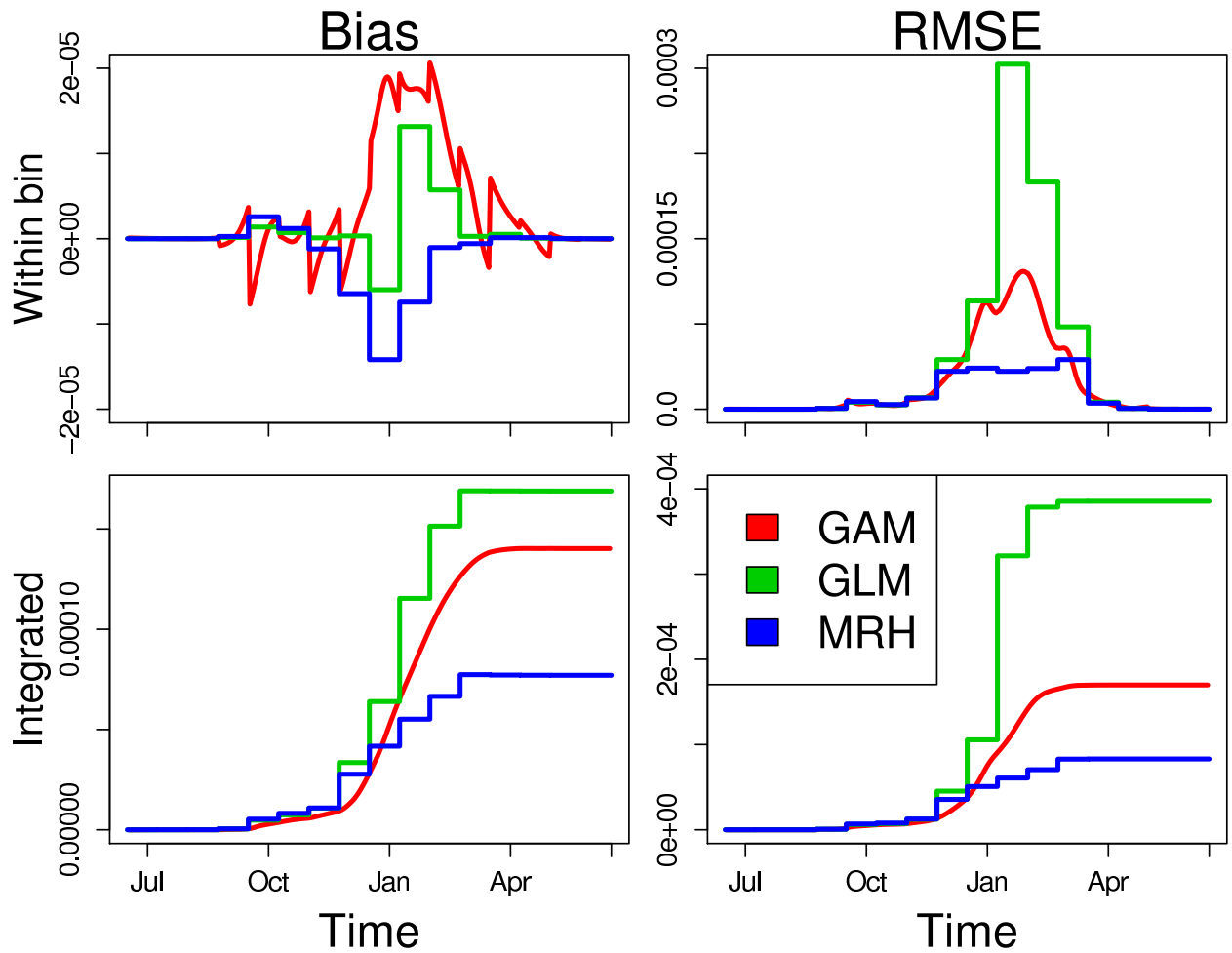


Figure 3: The bias and RMSE (calculated using equations (10) and (11)) over time for the hazard rate estimates from the 200 simulated data sets for the three models. The top two graphs contain the bias and RMSE calculated within each bin, and show that the MRH model tends to underestimate the hazard rate, while the Poisson GAM and Poisson GLM tend to overestimate the hazard rate. The MRH model has a smaller absolute bias than the other two. The RMSE calculations show that the MRH model has the smallest RMSE across the entire study period, followed by the Poisson GAM and Poisson GLM. The integrated bias and RMSE (bottom graphs) tell a similar story, with the smallest integrated bias and RMSE values for the MRH model, the largest integrated bias for the Poisson GAM, and the largest integrated RMSE for the Poisson GLM.

Table 1: Sample characteristics of the 11 years of data gathered in Navrongo, Ghana, including summaries of the monthly weather averages and percentages and monthly meningitis counts.

Variable	Mean	Median	SD	Range
Monthly Meningitis Cases	2.8	0.0	8.3	(0.0, 42.0)
Maximum Daily Temperature (Celsius)	35.4	36.0	3.1	(29.9, 41.1)
Minimum Daily Temperature (Celsius)	23.0	22.9	2.2	(18.1, 28.3)
Percent Relative Humidity (at 3pm)	39.5	36.5	20.6	(6.0, 73.0)
Percentage dusty days	36.8	0.0	44.1	(0.0, 100.0)
CO emissions (grams CO/day)	16.5	0.4	33.5	(0.0, 160.4)
Monthly Pneumonia Cases	1.4	0.0	2.4	(0.0, 14.0)

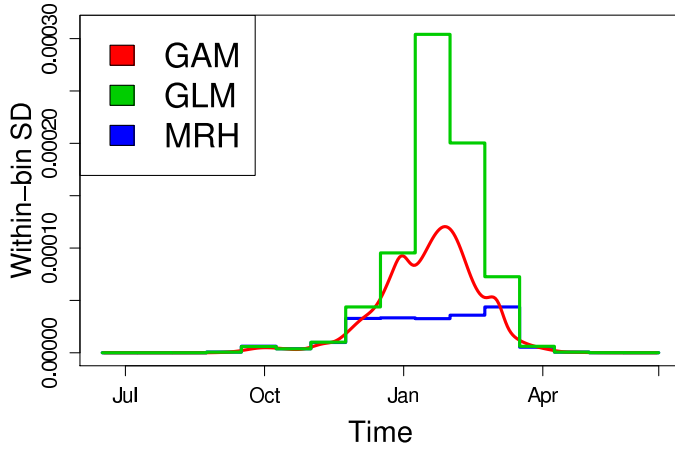


Figure 4: The within-bin standard deviation of each of the 200 estimates for each of the Poisson GLM, MRH model, and Poisson GAM. As shown in Section 3, the MRH model has a smaller variance than the Poisson GLM. From the simulations, it appears that the standard deviation of the estimator for the Poisson GAM is in between the other two.

poisson link in R. The Poisson GAM was implemented using `gam()` in the `mgcv` library in R, with the smoothing performed using cyclic cubic regression splines (penalized cubic regression splines with ends that match the beginning and end of the season) and basis dimension equal to 12. Offsets for both models are included as described in Sections 2.1 and 2.3.

To facilitate the dyadic splitting of the MRH model, the yearly period was divided into 16 subintervals. Since weather measures are only provided monthly, the measures were smoothed (using `smooth.spline()` in R) and then predicted (using `predict.smooth.spline()` in R) for the 16 mid-bin time points. Note that while this is a larger model compared to the other two models, there is little impact since the observations are aggregated monthly in the original data. The MRH model ran for 1 million MCMC iterations, with the first 250,000 burned, and the remaining thinned by 100 to reduce autocorrelation. The mean of the hyperpriors for a and b were fixed based on the Nelson-Aalen estimate of the cumulative hazard. The MRH model was implemented using an extension of the MRH package available in R[24]. For all models, the “year” was a seasonal year, and ranged from July to June so that meningitis outbreaks, which generally occur December to May were mid-season, avoiding smoothing issues near the boundary and enabling a continuous estimate of the function of seasonal meningitis trends.

The estimated hazard rate and covariate effects can be observed in Figure 5, which shows the average yearly hazard rate (top left), the 95% error bounds (95% credible intervals for the MRH model, and 95% confidence intervals for the Poisson GLM and Poisson GAM, top right), and the estimated effects of the time-varying covariates (bottom). The figure shows similar estimates for the hazard rate for the Poisson GLM and MRH model. However, the Poisson GAM shows a much smaller hazard rate than the other two, although still displays the mid-season meningitis peak. The 95% error bounds also seem similar among the models, although it is slightly difficult to compare the width of the GAM intervals with the other two models as it is smaller. The bottom graph of the figure shows almost identical estimates and 95% error bounds for all three models for the estimated effects of Dust, CO, and pneumonia, and minor differences with humidity. However, the estimates for the effects of minimum and maximum daily temperature show some major differences. While the Poisson GLM and MRH models show positive estimated effects for both temperature covariates, the MRH model shows a larger effect for the maximum daily temperature and a smaller effect for the minimum daily temperature, while the Poisson GLM shows the reverse. The Poisson GAM shows negative estimated effects for both temperature covariates, which explains the differences in the estimate of the hazard rate. These discrepancies are likely due to the collinearity between the two temperature variables.

The fitted monthly meningitis counts, plotted against the observed monthly meningitis counts, are shown in Figure 6, where $\hat{Y}_{ij} = W_{ij}\hat{\lambda}_{0j} \exp(X'_{ij}\hat{\beta} + Z'_{ij}\hat{\alpha})$. While all three models perform

similarly, the Poisson GLM and MRH model seem to do a better job capturing the true peaks than the Poisson GAM, which seems to overestimate the number of cases when the outbreaks are larger (such as in years 2000 and 2001). However, in years with very few meningitis counts, the Poisson GAM seems to be more reliable in capturing this than the other two models (such as in years 1999 and 2006). The difference between the observed and fitted values can be seen in Figure 7, which shows the monthly difference (top left), the square root of the monthly difference, squared (i.e.

$\sqrt{(Y_{ij} - \hat{Y}_{ij})^2}$, top right), the integrated absolute values of the monthly differences (bottom left), and the square root of the integrated square of the difference (bottom right). The integrated values have been multiplied by their bin width. It is hard to discern differences between the MRH model and Poisson GLM based on the within-bin residuals (top graphs), but it does seem more obvious that the Poisson GAM shows larger residuals than the other two models. The cumulative graphs (bottom) show the smallest residual differences for the MRH model, and equal squared differences between the MRH model and Poisson GLM. As shown in the top graphs, the Poisson GAM has the largest residuals of the three models.

6 Discussion

In this manuscript, we compare differences among the Poisson GLM, the MRH model, and the Poisson GAM. While the models are all very similar in the likelihood formulation and set-up, we observed differences (some minor, some major) in their performance. Because of its Bayesian structure and method of estimation, the MRH model provided the most accurate results in both the simulations and the analysis of the Navrongo data. In addition, we showed both theoretically and through simulations that the MRH estimator has the smallest standard deviation. However, the MRH model is more complex, and, as with most MCMC estimation routines, requires a longer run-time than the other two classical models. The simulations show that the Poisson GAM, with a smooth function to estimate the hazard rate, performs better than the Poisson GLM. However, in the real data analysis, the Poisson GLM had smaller residuals and did a better job predicting the meningitis counts over the 11-year period when compared to the Poisson GAM.

In the analysis of the Navrongo data, we are limited by the grouped data, and are unable to explore patterns in the contraction of meningitis at the subject-level. Two important factors that cannot be accounted for are the location of the meningitis case (at the subject-level), as those living in the city likely have different risks than those living in the more rural parts of Navrongo, as well as seasonal migration patterns (which affect the estimated number at risk). In addition, our analyses assume that the risk of contracting meningitis starts anew each season, and we cannot remove those who contracted meningitis from future cases. While the effect of these assumptions is likely small, it is possible that some of the cases recorded are for the same individual and that these individuals are at higher risk than others. Lastly, those who died from meningitis are also not recorded, and so may also be counted twice (once as a meningitis case, and then once as a censored subject), although this number is small and probably does not have much impact on the final results.

All models are sufficient in describing how weather and other predictors affect contraction of meningitis in the Sahel, and the choice of the best model should be based on the number of observed failures as well as time constraints for estimation.

Acknowledgements

This work was supported by grants nsf-geo 1211668, nsf-deb 1316334, and NIH-R01GM096655. The authors thank the researchers at NCAR for their helpful advice and input.

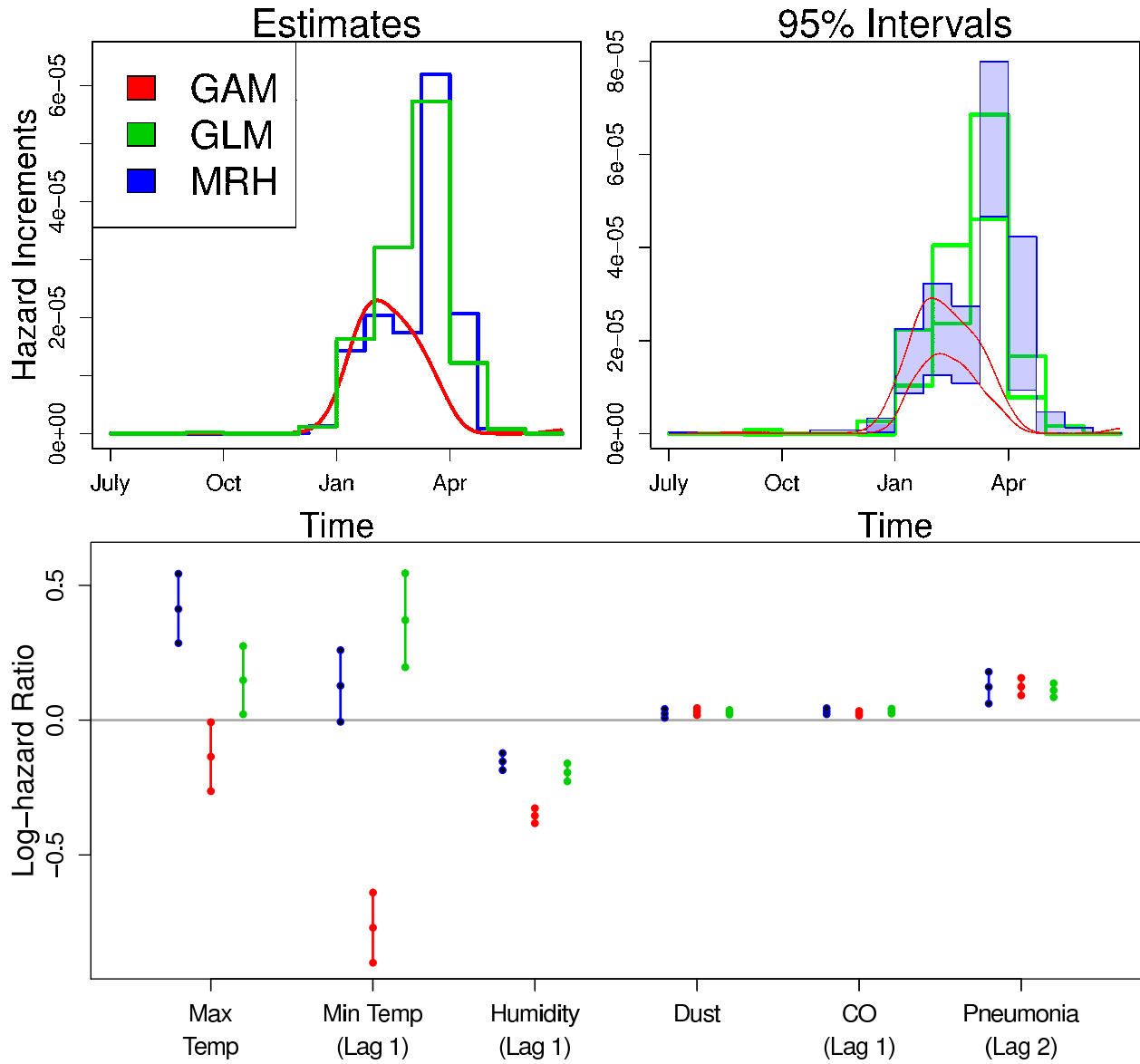


Figure 5: TOP: The estimated hazard rate of contracting meningitis (top left) and the 95% interval bounds (top right, with credible intervals for the MRH model and confidence intervals for the GLM and GAM Poisson models). While the MRH model hazard rate estimate differs slightly from the near-identical estimates of the Poisson GLM and GAM (left), the shape of all three models is similar: An increased risk of meningitis between January and April, with a peak in March, and virtually zero risk at other times of the season. On the right, the 95% interval bounds are shown. While the Poisson GAM and MRH models have intervals with similar widths and shapes, the GLM model is dramatically (and unsurprisingly) larger across the entire peak of the season. In addition, the bounds of the GLM go below zero, unlike the other two models.

BOTTOM: The estimated effects of the weather covariates on the meningitis risk. As expected, in all three models, increased daily temperatures (both minimum and maximum), percentage of dusty days, CO, and pneumonia cases increased the risk of meningitis. Higher humidity, on the other hand, decreased the risk of meningitis. All three models produced similar estimates and 95% interval bounds (credible intervals for the MRH model, confidence intervals for the Poisson GLM and GAM), with the exception of the maximum and minimum daily temperatures. The Poisson GLM and GAM are in agreement, with increases in minimum daily temperature having a larger effect on the increased meningitis risk than the maximum daily temperature. However, the MRH model shows the opposite, with increases in the maximum daily temperature having a larger effect on the increased meningitis risk. This is likely due to the fact that the two temperature covariates are somewhat collinear.

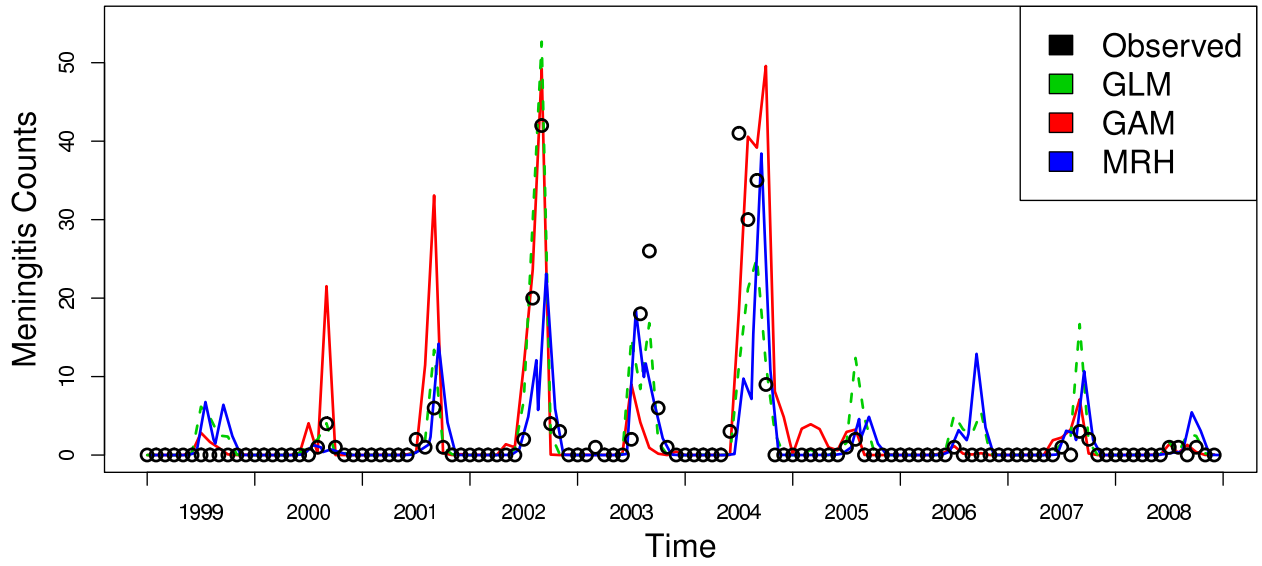


Figure 6: The estimated number of meningitis counts (i.e. the fitted values) over the 10-year seasonal period (July 1998 to July 2008), with the observed counts shown in black and the other models shown for comparison. The Poisson GAM and GLM produce virtually identical estimates, while the MRH model differs slightly. All three models tend to overestimate the small outbreaks (such as that in 1999, where there were zero meningitis confirmed meningitis cases, but all three models predicted at least a few). The Poisson GLM and GAM accurately captured the 2002 outbreak, while the MRH model underestimated the number of cases. However, in 2004, the MRH model accurately captured the high number of cases, while the other two models did not perform as well.

Appendix A: Full conditionals for the hyperparameters a, b, k , and $\gamma_{m,p}$

The parameters in the prior distributions of H and all $R_{m,p}$ s and a, b, k , and $\gamma_{m,p}$, can either be fixed at desired values, or treated as random variables with their own set of hyperpriors. Below are the forms of these full conditional distributions for a specific set of hyperpriors we chose. The notation η^- will be used to denote the set of all data and all parameters except for the parameter η itself. The full conditionals are as follows:

- If a is given a zero-truncated Poisson prior, $\frac{e^{-\mu_a} \mu_a^a}{a! (1 - e^{-\mu_a})}$ (chosen for computational convenience), the full conditional distribution for a is:

$$\pi(a | a^-) \propto \frac{H^a \mu_a^a}{b^a (a-1)! a!} \prod_{m=1}^M \prod_{p=0}^{2^{m-1}-1} \left\{ \frac{R_{m,p}^{2\gamma_{m,p} k^m a} (1 - R_{m,p})^{2(1-\gamma_{m,p}) k^m a}}{B(2\gamma_{m,p} k^m a, 2(1 - \gamma_{m,p}) k^m a)} \right\}$$

- If the scale parameter b in the gamma prior for the cumulative hazard function H is given an exponential prior with mean μ_b , the resulting full conditional is:

$$\pi(b | b^-) \propto \frac{1}{b^a} \exp \left\{ - \left(\frac{H}{b} + \frac{b}{\mu_b} \right) \right\}$$

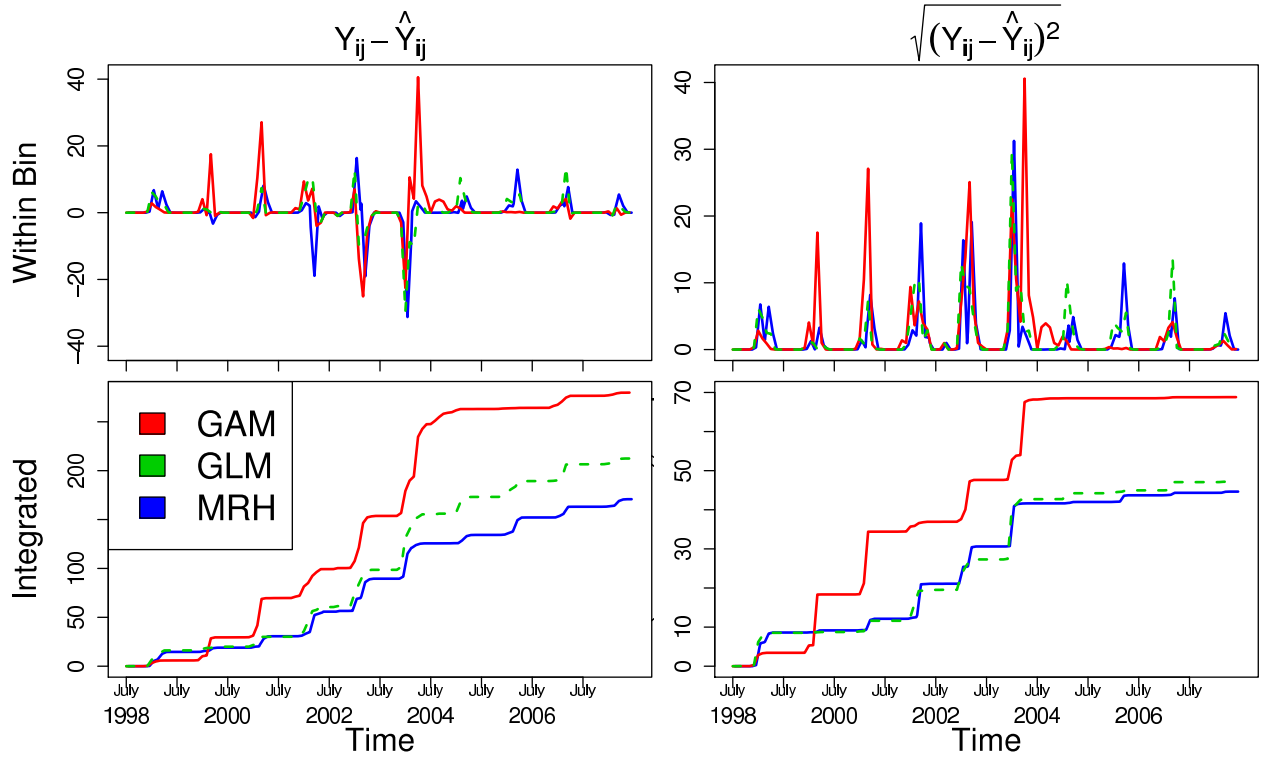


Figure 7: Differences between the observed and fitted values for the three models. The within-bin values of the difference (top left), and the square root of the squared difference (top left) show only a little variation between the three models. However, the integrated absolute difference (bottom left) and integrated squared difference (bottom right) show obvious, smaller integrated values for the Poisson MRH model. The integrated values are the cumulative differences, multiplied by the bin width (equal to one for the Poisson GLM and GAM, and equal to 12/16 for the Poisson MRH).

- If k is given an exponential prior distribution with mean μ_k , the full conditional distribution for k is as follows:

$$\pi(k \mid k^-) \propto \prod_{m=1}^M \prod_{p=0}^{2^m-1} \left\{ \frac{R_{m,p}^{2\gamma_{m,p}k^ma} (1 - R_{m,p})^{2(1-\gamma_{m,p})k^ma}}{B(2\gamma_{m,p}k^ma, 2(1-\gamma_{m,p})k^ma)} \right\} e^{-\frac{k}{\mu_k}}$$

- If a Beta(u, w) prior is placed on each $\gamma_{m,p}$, the full conditional distribution for each $\gamma_{m,p}$ is proportional to:

$$\frac{R_{m,p}^{2\gamma_{m,p}k^ma} (1 - R_{m,p})^{2(1-\gamma_{m,p})k^ma}}{B(2\gamma_{m,p}k^ma, 2(1-\gamma_{m,p})k^ma)} \gamma_{m,p}^{u-1} (1 - \gamma_{m,p})^{w-1}$$

References

- [1] A. Abdussalam, A. Monaghan, V. Dukić, M. Hayden, T. Thomson, G. Leckebusch, and J. Thornes. Climate influences on meningitis incidence in northwest Nigeria. *Weather, Climate, and Society*, 6:62–76, 2014.
- [2] L. Agier, M. Stanton, G. Soga, and P. Diggle. A multi-state spatio-temporal Markov model for categorized incidence of meningitis in sub-Saharan Africa. *Epidemiology and infection*, 141:1764–1771, 2013.
- [3] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:716–723, 1974.
- [4] A Beresniak, E Bertherat, W Perea, G Soga, R Souley, D Dupont, and S Hugonnet. A Bayesian network approach to the study of historical epidemiological databases: modelling meningitis outbreaks in the Niger. *Bulletin of the World Health Organization*, 90(6):412–417a, 2012.
- [5] J. Besancenot, M. Boko, and P. Oke. Weather conditions and cerebrospinal meningitis in Benin (Gulf of Guinea, West Africa). *European Journal of Epidemiology*, 13:807–815, 1997.
- [6] P. Bouman, J. Dignam, V. Dukic, and X.L. Meng. A multiresolution hazard model for multi-center survival studies: Application to Tamoxifen treatment in early stage breast cancer. *Journal of the American Statistical Association*, 102:1145–1157, 2007.
- [7] P. Bouman, V. Dukic, and X.L. Meng. Bayesian multiresolution hazard model with application to an AIDS reporting delay study. *Statistica Sinica*, 15:325–357, 2005.
- [8] H. Broutin, S. Philippon, G. Constantin de Magny, M. Courel, B. Sultan, and J. Guégan. Comparative study of meningitis dynamics across nine Arican countries: a global perspective. *International Journal of Health Geographics*, 6:29–37, 2007.
- [9] J. Cheesbrough, A. Morse, and S. Green. Meningococcal meningitis and carriage in western Zairea hypoendemic zone related to climate. *Epidemiology and Infection*, 114:75–92, 1995.
- [10] Y. Chen, Y. Hagar, J. Dignam, and V. Dukic. Pruned Multiresolution Hazard (PMRH) models for time-to-event data. *Bayesian Analysis*, In Review, 2014.
- [11] L. Cuevas, I. Jeanne, A. Molesworth, M. Bell, E. Savory, S. Connor, and M. Thomson. Risk mapping and early warning systems for the control of meningitis in Africa. *Vaccine*, 5:A12–A17, 2007.

- [12] P. Diggle, Y. Guan, A. Hart, F. Paize, and M. Stanton. Estimating individual-level risk in spatial epidemiology using spatially aggregated information on the population at risk. *Journal of the American Statistical Association*, 105(492), 2010.
- [13] J. Dignam, V. Dukic, S. Anderson, E. Mamounas, D. Wickerham, and N. Wolmark. Hazard of recurrence and adjuvant treatment effects over time in lymph node-negative breast cancer. *Breast Cancer Research and Treatment*, 116:595–602, 2009.
- [14] J. Dignam, V. Dukic, S. Anderson, E. Mamounas, D. Wickerham, and N. Wolmark. Hazard of recurrence and adjuvant treatment effects over time in lymph node-negative breast cancer. *Breast Cancer Research and Treatment*, 116:595–602, 2009.
- [15] V. Dukic and J. Dignam. Bayesian hierarchical multiresolution hazard model for the study of time-dependent failure patterns in early stage breast cancer. *Bayesian Analysis*, 2:591–610, 2007.
- [16] V. Dukić, M. Hayden, T. Hopson A. Forgor, P. Akweongo, A. Hodgson, A. Monaghan, C. Wied-inmyer, T. Yoksas, T. Thomson, S. Trzaska, , and R. Pandya. The role of weather in meningitis outbreaks in Navrongo, Ghana: A generalized additive modeling approach. *Journal of Agricultural, Biological and Environmental Statistics*, 17:442–460, 2012.
- [17] T. Ferguson. Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2:615–629, 1974.
- [18] C.P. García-Pando, M. Stanton, P. Diggle, S. Trzaska, R. Miller, J. Perlwitz, J. Baldasano, E. Cuevas, P. Ceccato, P. Yaka, and M. Thomson. Soil dust aerosols and wind as predictors of seasonal meningitis incidence in Niger. *Environmental health perspectives*, 122:679–686, 2014.
- [19] M. Glasser. Exponential survival with covariance. *Journal of the American Statistical Association*, 62:561–568, 1967.
- [20] B. Greenwood. Meningococcal meningitis in Africa. *Transaction of the Royal Society of Tropical Medicine and Hygiene*, 93:341–353, 1999.
- [21] B. Greenwood, I. Blakebrough, A. Bradley, S. Wali, and H. Whittle. Meningococcal disease and season in sub-Saharan Africa. *Lancet*, 1:1339–1342, 1984.
- [22] B. Greenwood, A. Greenwood, A. Bradley, K. Williams, M. Hassan-King, F. Shenton, R. Wall, and R. Hayes. Factors influencing susceptibility to meningococcal disease during an epidemic in The Gambia, West Africa. *Journal of Infection*, 14:167–184, 1987.
- [23] Y. Hagar, D. Albers, R. Pivavarov, H. Chase, V. Dukic, and N. Elhadad. Survival analysis with Electronic Health Record data: Experiments with Chronic Kidney Disease. *Statistical Analysis and Data Mining*, 7:385–403, 2014.
- [24] Y. Hagar, Y. Chen, and V. Dukic. MRH package in R. <http://cran.r-project.org/web/packages/MRH/index.html>, 2014.
- [25] T. Hanson. Inference for mixtures of finite Polya tree models. *Journal of the American Statistical Association*, 101:1548–1565, 2006.
- [26] T. Hanson and W. Johnson. Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association*, 97:1020–1033, 2002.

- [27] T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1:297–318, 1986.
- [28] T. Holford. The analysis of rates and of survivorship using log-linear models. *Biometrics*, 36:299–305, 1980.
- [29] T.R. Holford. Life tables with concomitant information. *Biometrics*, 32:587–597, 1976.
- [30] JG. Ibrahim, M. Chen, and D. Sinha. *Bayesian Survival Analysis*. Springer, New York, 2001.
- [31] J.D. Kalbfleisch and R.L. Prentice. *The Statistical Analysis of Failure Time Data*. SWiley, Chichester, 2002.
- [32] J. Klein and M. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag, New York, New York, USA, 2003.
- [33] N. Laird and D. Olivier. Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76:231–240, 1981.
- [34] L. Lapeyssonnie. Cerebrospinal meningitis in Africa. *Bulletin of the World Health Organization*, 28:3–114, 1963.
- [35] P. Laplace. Memoir on the probability of causes and events. *Statistical Science*, 6:364–378, 1986.
- [36] M. Lavine. Some aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, 20:1222–1235, 1992.
- [37] J. Leimkugel, A. Hodgson, A. Adams Forgor, V. Plugger, J. Dangy, T. Smith, M. Achtman, S. Gagneux, and G. Pluschke. Clonal waves of *Neisseria* colonisation and disease in the African meningitis belt: Eight-year longitudinal study in northern Ghana. *PLoS Medicine*, 4:535–544, 2007.
- [38] A. Molesworth, L. Cuevas, S. Connor, A. Morse, and M. Thomson. Environmental risk and meningitis epidemics in Africa. *Emerging Infectious Diseases*, 9:1287–1293, 2003.
- [39] A. Molesworth, L. Cuevas, and M. Thomson. Forecasting meningitis epidemics in Africa. *Liverpool: LSTM*, 2002.
- [40] P. Moore. Meningococcal meningitis in sub-Saharan Africa: A model for the epidemic process. *Clinical Infectious Diseases*, 14:515–525, 1992.
- [41] P. Moore, M. Reeves, B. Schwartz, B. Gellin, and C. Broome. Intercontinental spread of an epidemic group “*A Neisseria Meningitidis* strain”. *Lancet*, 2:260–263, 1989.
- [42] P. Muliere and S. Walker. A Bbayesian non-parametric approach to survival analysis using Polya trees. *Scandinavian Journal of Statistics*, 24:331–340, 1997.
- [43] J. Nelder and R. Baker. Generalized linear models. *Encyclopedia of Statistical Sciences*, 1972.
- [44] L. Nieto-Barajas and P. Müller. Rubbery Polya tree. *Scandinavian Journal of Statistics*, 39:166–184, 2012.
- [45] P. Nyarko, P. Wontuo, A. Nazzar, J. Phillips, P. Ngom, and R. Binka. *Navrongo DSS Ghana*. Navrongo DSS Ghana, Ministry of Health, Navrongo Health Research Centre, Navrongo, Ghana, 2000.

- [46] A. Oduro, G. Wa, D. Azongo, C. Debpuur, P. Wontuo, F. Kondayire, P. Welaga, A. Bawah, A. Nazzar, J. Williams, A. Hodgson, and F. Binka. Profile of the Navrongo health and demographic surveillance system. *International Journal of Epidemiology*, 41:968–976, 2012.
- [47] R.L. Prentice. Exponential survivals with censoring and explanatory variables. *Biometrika*, 60:279–288, 1973.
- [48] L. Roberts. An ill wind, bringing meningitis. *Science*, 320:1710–1715, 2008.
- [49] M. Stanton, L. Agier, B. Taylor, and P. Diggle. Towards realtime spatiotemporal prediction of district level meningitis incidence in sub-Saharan Africa. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 177:661–678, 2014.
- [50] B. Sultan. Influence of climate upon the meningitis onset in West Africa. *Medicine Sciences*, 21:470–471, 2005.
- [51] B. Sultan, L. Chiapello, and M. Aouam. Le rôle du climat et des aérosols sur les Épidémies de méningite en Afrique de l’Ouest. *PLoS Medicine*, 2:43–49, 2005.
- [52] B. Sultan, K. Labadi, J. Guégan, and S. Janicot. Climate drives the meningitis epidemics onset in West Africa. *PLoS Medicine*, 2:43–49, 2005.
- [53] M. Thomson, A. Molesworth, M. Djingarey, K. Yameogo, F. Belanger, and L. Cuevas. Potential of environmental models to predict meningitis epidemics in Africa. *Tropical Medicine and International Health*, 1:781–788, 2006.
- [54] C. Wiedinmyer, S. Akagi, R. Yokelson, L. Emmons, J. Al-Saadi, J. Orlando, and A. Soja. The Fire Inventory from NCAR (FINN) – a high resolution global model to estimate the emissions from open burning. *Geoscientific Model Development Discussions*, 3:2439–2476, 2011.
- [55] W. Wong and L. Ma. Optional Polya tree and Bayesian inference. *The Annals of Statistics*, 38:1433–1459, 2010.
- [56] P. Yaka, B. Sultan, H. Broutin, S. Janicot, S. Philippon, and N. Fourquet. Relationships between climate and year-to-year variability in meningitis outbreaks: A case study in Burkina Faso and Niger. *International Journal of Health Geographics*, 7:34, 2008.
- [57] L. Zhao and T. Hanson. Spatially dependent Polya tree modeling for survival data. *Biometrics*, 67:391–403, 2011.
- [58] L. Zhao, T. Hanson, and B. Carlin. Mixtures of Polya trees for flexible spatial frailty survival modelling. *Biometrika*, 96:263–276, 2009.